Confidence Measure Estimation in Dynamical Systems Model Input Set Selection

Paul B. Deignan, Jr.
Galen B. King
Peter H. Meckl

School of Mechanical Engineering
Purdue University
West Lafayette, IN 47907-1288

Kristofer Jennings

Department of Statistics
Purdue University
West Lafayette, IN 47907-2067

Abstract

An information-theoretic input selection method for dynamical system modeling is presented that qualifies the rejection of irrelevant inputs from a candidate input set with an estimate of a measure of confidence given only finite data. To this end, we introduce a method of determining the spatial interval of dependency in the context of the modeling problem for bootstrap mutual information estimates on dependent time-series. Additionally, details are presented for determining an optimal binning interval for histogram-based mutual information estimates.

Introduction

It is apparent that continued improvement in practical nonlinear control theory relies on the ability to formulate accurate models of system dynamics. As we strive to control the dynamics of increasingly complex and varied systems, we are confronted with the choice of whether to be satisfied with a controller that performs reliably but not optimally or to attempt to construct a model that is characteristic of the essential system dynamics and accurate so that our controller design problem is not so constrained. Ultimately, we are forced to consider how to deal with uncertainties in the modeling process. Our purpose here is to provide a method for determining the appropriate input space in which empirically derived models of dynamical systems should be constructed given a quantifiable degree of confidence for considering an input dimension relevant.

Whether the system is to be described by a map or a dynamical manifold, the input space is chosen from a set of measurable, related variables so that the trajectory of the variable or variables of interest lies in a coherent subspace of small enough dimensionality as to provide good estimates with limited data. Locally, the dynamical manifold is a coordinate map in Euclidean space that defines a functional relationship between inputs and an output as in the prototypical system identification problem. Globally, however, the manifold defines a multi-valued relation between variables so a measure of central tendency is inappropriate as a criterion of merit for the global input selection problem; a more suitable measure of dependence is mutual information (MI). The determination of the optimal input set of finite dimension is a problem in combinatorial optimization. The most effective means of decreasing the computational challenge of this problem is to cull the candidate set of inputs that have no potential of membership in the final optimal set. To do this with a measure of confidence for arbitrary distributions, we rely on the bootstrap estimate of the criterion of merit. However, since data describing a deterministic dynamical system is obviously not independent and identically distributed (i.i.d.), we must first determine the spatial interval of dependence in order to perform the bootstrap.

After presenting pertinent background information regarding the estimation of MI using multi-dimensional, histograms based on uniform bin widths, we show how an optimal binning may be selected given one weighting parameter between histogram smoothness and estimated mutual information. This result is used to determine a spatial stratification interval of local dependence so that a bootstrap estimate of mutual information can be calculated. Finally, we demonstrate how individual candidate input dimensions can be systematically eliminated from the candidate pool based on the MI measure of deterministic relevance to an empirical dynamical manifold model of the system generating the single observed output. Results of this method of input selection are demonstrated on a data set derived from measurements of diesel engine operation with net engine torque as the output signal.

Input Selection by Mutual Information

This section provides a method of estimation of mutual information as a measure of joint nonlinear dependence. Histograms based on uniformly binning the system observations over the range of values are chosen rather than kernels to approximate the nonparametric probability density function due to their computational efficiency and the insignificance of relative accuracy of the resultant probability density function (pdf) estimates in sparsely populated spaces. Therefore, we will first present a method to systematically arrive at an optimal uniform binning interval for the estimation of joint mutual information.

Mutual Information

When restricted to finite probability spaces, normalized measures of mutual information satisfy all of the Rényi postulates for measures of dependence (Bell C.B., 1962). Given this property and the result that mutual information may be estimated to an arbitrary degree of resolution uniformly in a measurement continuum, we will consider mutual information to be an optimal measure of dependence. In a measurement continuum it is possible to take an unlimited number of measurements with infinite precision.

It is not necessary to normalize the measure for the purpose of feature subset selection. Mutual information quantifies the uncertainty in the system output, Y, that is conditioned on the system input, **X**. Specifically, average mutual information between two signals is the Kullback-Leibler distance, $I(\mathbf{X};Y)$ between the conditional probability distribution, $p(\mathrm{Y} \mid \mathbf{X})$, and the individual output probability distribution, $p(\mathrm{Y})$, and is given by the formula

$$I(\mathbf{X};Y) = \sum_{\mathbf{x}\in\aleph} \sum_{y\in\Im} p(\mathbf{x},y) \log \frac{p(y \mid \mathbf{x})}{p(y)},$$

where **X** is the input vector and Y is the output and the summation is over the discrete values **x** and y of the random variables **X** and Y, respectively. If the natural logarithm is used, the average mutual information is in natural units (nats). This formulation represents a difference in entropy of the data distributions,

$$I(\mathbf{X};\mathrm{Y}) = H(\mathrm{Y}) - H(\mathrm{Y} \mid \mathbf{X}),$$

where entropy, $H(\cdot)$, is defined by the relation:

$$H(\mathbf{X}) = -\sum_{\mathbf{x}\in\aleph} p(\mathbf{x}) \log(p(\mathbf{x})).$$

Entropy represents the uncertainty in each signal. Thus, average mutual information represents the decrease in uncertainty of the output, *y*, which may be achieved through the knowledge of the input, **x**, alone. The average mutual information is zero if the output data distribution is invariant to the occurrence of the particular input vector, *i.e.*, $H(\mathrm{Y} \mid \mathbf{X}) = H(\mathrm{Y})$ (Cover and Thomas, 1991).

Average mutual information may be more easily calculated by applying the Bayes' Rule to the conditional probability, i.e.,

$$\begin{aligned} I(\mathbf{X};Y) &= \sum_{\mathbf{x}\in\aleph} \sum_{y\in\Im} \hat{p}(\mathbf{x},y) \log \frac{\hat{p}(\mathbf{x},y)}{\hat{p}(\mathbf{x})\hat{p}(y)} \\ &= \sum_{\mathbf{x}\in\aleph} \sum_{y\in\Im} \hat{p}(\mathbf{x},y) \left(\log \hat{p}(\mathbf{x},y) - \log \hat{p}(\mathbf{x}) - \log \hat{p}(y) \right) \end{aligned}$$

When performing the estimates for a fixed record size, *N* with the binning frequencies given by *n*, the equation is calculated as

$$I(\mathbf{X};Y) = \log(N) + \frac{1}{N} \sum_{\mathbf{x}\in\aleph} \sum_{y\in\Im} n_{\mathbf{x},y} \left(\log n_{\mathbf{x},y} - \log n_{\mathbf{x}} - \log n_{y} \right)$$

which may be a root-n consistent estimator of self-information (*Y* in place of **X**) for some binwidth under a mild assumption of bounded tail behavior (Hall and Morton, 1993). Note that the measure can be easily aggregated for multiple scopes when probability estimates are calculated by uniformly binning the data. Uniform binning over the span of the data values also yields estimates that are invariant to affine transformations, such as those caused by improperly calibrated linear sensors.

The entropy, $H(Y)$, determines the maximum information content of the output signal. This quantity provides an upper limit on the achievable mutual information content of any input signal set given the structure of the output data and the binning used to estimate the joint and individual probabilities.

Uniform Binning for MI Estimation

The choice of binning interval for mutual information estimation is a compromise between resolution and computing resources within limits established by the underlying probability distribution and the practical availability of data. The interval should be chosen so that relevant structure of the joint distributions is resolved while allowing joint mutual information estimates to be computed up to the dimensionality of the feature set. Because we are ultimately interested in the relative rather than absolute contribution of the various inputs to the accuracy of the mapping, the binning interval of the base distribution is held fixed while comparing the incremental increase of estimated joint mutual information with each additional candidate input. For pairwise mutual information comparisons, this means that the binning interval of the output signal is held fixed at the optimal binning interval of the output entropy.

The goal of binning is to reveal structure in a probability distribution while minimizing the unavoidable effect of quantization of a signal that is essentially continuous into a number of bins, the heights of which are integer multiples of $1/N$, where *N* is the number of data records. Mutual information is a convex functional of continuous probability density functions. When estimated by relative frequency, mutual information is nondecreasing with number of binning partitions. In the limit for infinite data, the rate of estimated MI increase is inverse log-linear in the absence of any informational structure in the data. This is a fundamental property of the estimate as used in this analysis. We present this formally in the following lemma:

Lemma 1: With probabilities estimated by relative frequencies, the mutual information functional, $I(\mathbf{X};Y)$, is nondecreasing with the number of partitions for random variables $\boldsymbol{X}$ and Y and constant with partition number if and only if the events are uniformly distributed.

Proof

Let n be the number of occurrences of an event. Estimating the pdf consistently by relative frequency with one bin, the estimated mutual information is

$$I_1(\mathbf{X};Y) = -\frac{1}{n}\log\left(\frac{1}{n}\right).$$

Subdividing this bin into n_1 and n_2 events, we have that

$$I_2(\mathbf{X};Y) = -\left(\frac{1}{n_1}\log\left(\frac{1}{n_1}\right)+\frac{1}{n_2}\log\left(\frac{1}{n_2}\right)\right)$$

where $n = n_1 + n_2$.

It follows immediately from Jensen's discrete inequality,

$$f\left(\sum_{i=1}^{N} p(\boldsymbol{x}_i)\right) \le \sum_{i=1}^{N} p\left(f(\boldsymbol{x}_i)\right)$$

for strictly convex functions, f (such as $\boldsymbol{x}\log\boldsymbol{x}$ for $\boldsymbol{x} \ge 0$) that

$$I_1(\mathbf{X};Y) \le I_2(\mathbf{X};Y)$$

with equality if and only if $n_1 = n_2$. The generalization of this relation for any countable number of partitions follows inductively by repeated application of the inequality. Now, if the division of bins is such that all bins are nonempty and occurrences equally distributed, then for a k-partition the total MI of all bins is estimated as

$$I_k(\mathbf{X};Y) = -\frac{k}{k}\log\left(\frac{1}{k}\right)$$
$$= \log(k).$$

So, in the absence of probabilistic structure, the estimated MI increases asymptotically inversely log-linear as a function of number of bins. ■

Note that the controlling assumption of Lemma 1 in the bins/MI relation is that the measurements are sufficiently dense so that on average the occurrences are nearly equally distributed. Naturally, this assumption is not valid in the case of finite data sets or coarsely quantized measurements beyond a certain binning interval. This limitation is shown quantitatively in the estimation of self-information of engine torque from a data set describing the operation of a diesel engine over a 20 minute test cycle (Figure 1) and qualitatively from histograms (Figures 2a-c).

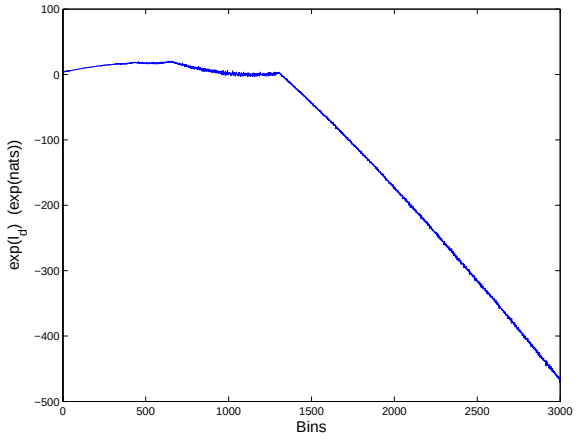

Figure 1. Exponential of torque self-information linearly detrended by first 1500 bins showing binning saturation effect for the data set of 6009 records.

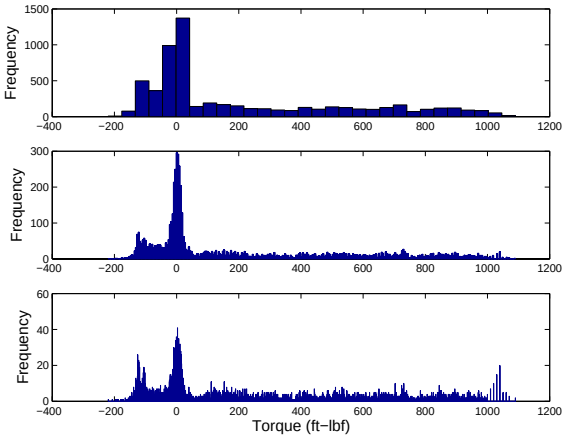

Figure 2a-c. Torque measurement distribution estimation by uniform binning **a)** 30 bins **b)** 300 bins **c)** 3000 bins

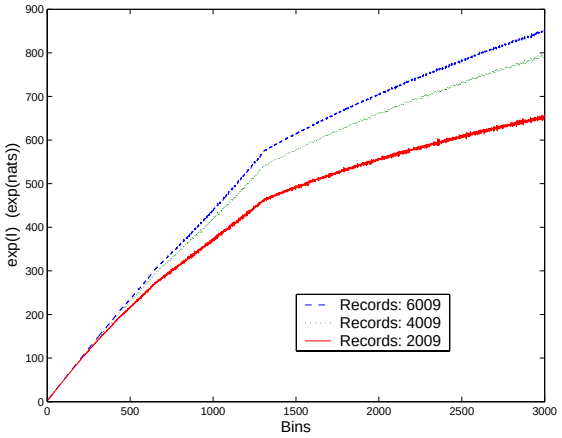

Figure 3. exp(I) v. bins trend per data set size of uniformly randomly decimated time-series torque records.

Probabilistic structure is revealed by the variability of the graph of MI over lower binning ranges. Unless the data is pathological, quantization noise will increase incrementally with higher binning resolutions yielding an asymptotically linear trend of the exponential of mutual information with number of bins. Figure 1 shows a structure region between 5 to 600 bins followed by a region where quantization noise due to the sampling resolution dominates and finally a region above 1400 bins where the dominant trend is attributable to the fact that the data record length is fixed. The estimates of MI are nondecreasing with number of bins; however the rate of increase will be less than that predicted by Lemma 1, giving a negative skew to the graph.

Since the MI estimate is generated from relative frequencies of measurement values, we expect the qualitative and quantitative properties of the estimate to be robust to random measurement variation. Figure 3 confirms this while also demonstrating the consistency of the saturation effect and the well-known bias in estimates due to data set size. Mutual information estimates are proportional to the average number of occurrences per binning interval. Furthermore, the nature of the trend is not significantly a function of binning interval in this example where the data is randomly decimated from 6009 records to 2009. It is interesting to note that the bias is downward rather than the upward bias predicted by the first order term of some expansions of the estimate (Treves and Panzeri, 1995).

We would like to choose a binning interval that is in some sense optimal for MI estimates. One strategy for determining an optimal bin width would be to search from a small number of bins upward while the skewness of the trend remains within a certain tolerance. This method is not generally practical as the smoothness—equivalently roughness—of the distribution is uncertain and probabilistic structure will potentially confuse the determination of the range. Rather than constructing a complicated set of statistical rules, it seems preferable to determine the feasible binning range by a more direct evaluation of the estimated MI per bins graph.

Here we are concerned with data derived from a regularly sampled time series of a continuous process so that there is a regularity condition that can be combined with a computational constraint in the penalty function, R, of the penalized objective function,

$$bin_{opt} = \arg\max\left(\hat{I} - \lambda R\right)$$

where λ is the penalty factor. Following Scott (1992), we chose R to be a bias corrected estimator derived from a finite difference approximation of a composition of the sampling and continuous process functions,

$$R = k^3\left(\sum_{i=1}^{k-1}\left(\frac{n_{i+1}-n_i}{N}\right)^2 - \frac{2}{N}\right).$$

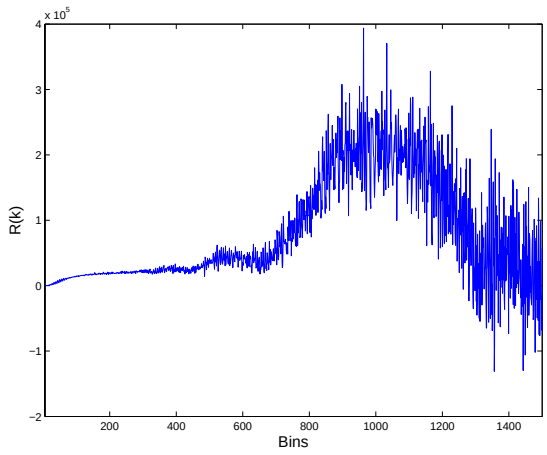

Figure 4. Roughness measure of torque distribution

This objective function brings together measures of per record data coherence in the form of the mutual information estimate as well as the coherence of the measured data by value through the roughness factor. Note that while the optimization function might be applied to categorical data with R serving as simply a computational constraint, the optimal bin width is more likely to be predetermined by the number of categories. Naturally, the search for the optimal binning interval should be limited to a binning range that exhibits probabilistic structure and is below that range where binning saturation is dominant.

Given a limited search interval determined by the user, our cost function is a more direct measure than that of Hall and Morton (1993) where the penalty factor is given as the data record length normalization of number of nonempty bins as it accounts for quantization noise due to sampling resolution as well. As a function of bins, the trend in the roughness measure corresponds with the qualitative observations of the torque distribution (Figures 2,4). The feasible range for the optimal bin search appears to be 5 to 600 bins. Above this, the trend departs radically.

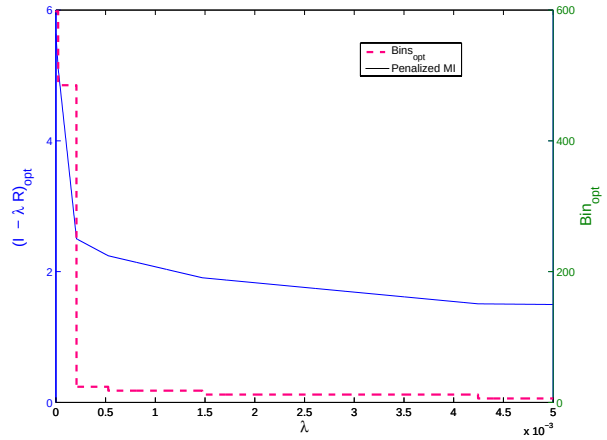

Figure 5. Optimal penalized self-information and binnning for the torque distribution

The self-information estimate of a record set is maximal when all occurrences are individually binned. Note that the mutual and self-information estimates are invariant to the spatial ordering of the bins. The weighting of the penalized MI estimate by roughness not only imposes a spatial ordering constraint on the distribution, it effectively transforms the binwidth variable into a sensitivity measure of average entropy per unit length. Thus, the optimal binwidth can be considered to be a measure of local dependence in one variable or, alternatively, as a quantization of self-information over that dimension. If in respect to this variable and binwidth there is a measurable dependence with a second variable, then within the region of the product space determined by the optimal binwidth of each dimension, the measurement pairs are dependent by the very same metric that is used to define dependence between variables. Generically, for the event pair (ω_1, ω_2),

$$P(\omega_1)P(\omega_2) \neq P(\omega_1, \omega_2)$$

Conservatively, the measurement pairs must be on average separated by at least one optimal binwidth in each respective dimension to be considered independent.

There is a substantial difference in the optimal binning over a small range of λ (0.0001-0.001) between which values the optimal binning is {485, 24} (see Figure 5). We would like to choose the binning corresponding to a λ value that yields the greatest estimated MI provided the distribution is as smooth as should be expected. Since binning above 400 is still rather suspect due to quantization noise (Figure 4), an optimal binning of 24 is chosen. Additional smoothing comes at the expense of useful MI. Our choice of weighting factor here is from a small set of feasible values; this is not expected to be the case in general. Ultimately, the choice is also a function of other factors such as *a priori* knowledge of the data source and characteristics, capacity of computational resources, the problem to be addressed by the analysis, and other factors that may be difficult to quantify as a cost function. The simplification of the only two competing terms is one way to easily incorporate these considerations in the input selection problem.

Model Input Selection Algorithm

It is not generally possible to find "root-n" consistent estimates of mutual information using histograms (Hall and Morton, 1993). However, this is beside the point as our purpose here is to determine relatively which inputs are optimal to a certain acceptable numerical resolution. The input selection problem is followed by a mapping problem which carries with it a degree of imprecision dependent on the model structure and independent of the particular inputs chosen. Therefore, the problem of input selection should be cast in terms of what is computationally discernible rather than what is theoretically possible. It is sufficient for our purposes that the manner in which we chose binning intervals for the various prospective inputs does not generate a bias in the evaluation procedure of making subsequent selections.

The input selection process for model formulation begins with the determination of the optimal binning of the output self-information. This interval is then fixed and the optimal binning interval of estimated mutual information is likewise determined for all input-output pairs and held constant in all joint MI comparisons. The greatest estimated joint MI input set is selected as the space in which to cast the model construction problem of a fixed dimension. Since joint estimates of MI are nondecreasing with dimensionality, the combinatorial input selection optimization process is amenable to branch and bounding algorithms which eliminates irrelevant inputs from the candidate input set systematically. However, to reject inputs as irrelevant normally requires the knowledge of a level of confidence at which a certain candidate input is irrelevant. Note that the dependency between input-output pairs is not the same as the dependency between samples.

Bootstrapping Dependent Data for Model Formulation

The bootstrap method of estimating confidence intervals, which is applicable to functionals of general probability distribution functions relies on i.i.d. resampling (Efron, 1979). By hypothesis, the data is generated from a deterministic process of unknown dimensionality, but with spatially local dependencies invariant with time. As a result of the model input selection algorithm, the binning intervals relative to a maximally informational output interval have been determined. The interval of spatial dependency of the output may be taken as twice the binning interval of the output entropy. Likewise, the input interval of spatial dependency is taken as twice that of the optimal binning interval of the mutual information estimate of that input. The upper bound of joint MI of a branch of the combinatorial search of inputs may be estimated as a summation of pairwise estimates with correction (Deignan, et *al*., 2003). The confidence measure of rejection is obtained through a composition of distributions of pairwise MI bootstrap estimates.

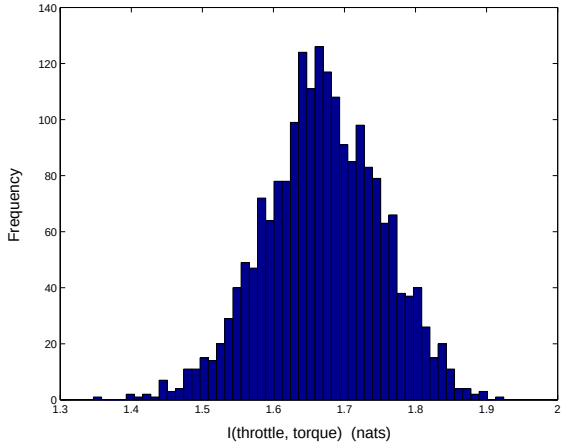

Figure 6. Histogram of bootstrap estimate of I(throttle,torque) at Bins: (27, 24), 2000 estimates

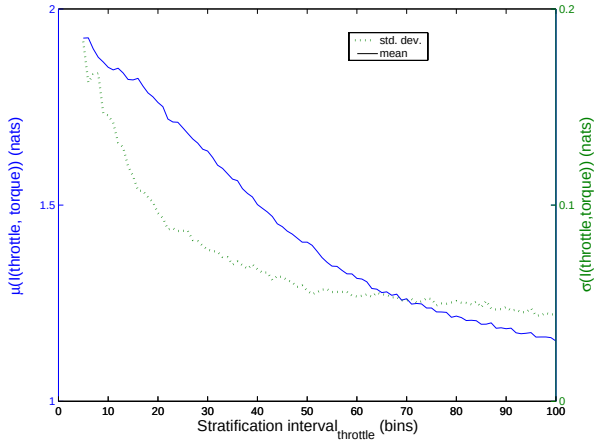

Figure 7. Statistical trends of bootstrap estimate of I(throttle,torque) by width of throttle stratification interval.

The uncertainty in the estimate of MI relevant to input rejection is that which is a function of the uncertainty inherent in the data alone and not a function of the determinism of the underlying system. Sampling to generate the analog of the distribution is typically done by a Monte Carlo scheme with replacement (Shao and Tu, 1995). Overlaying this we adopt a stratified sampling scheme without replacement based on the local spatial input-output dependency to produce the final sample on which the bootstrap estimate is calculated. Strata are thus defined as a rectangular region twice the width of the respective optimal bin width in that dimension. In the case of our data, this resampling scheme produces an approximately Gaussian distribution of bootstrap estimates of *I*(*throttle*,*torque*) (Figure 6). The choice of strata is verified for the throttle dimension by sweeping the bootstrap estimates as a function of stratification interval which shows that the standard deviation is approximately constant over twice the optimal bin width of 27 thereby indicating independence of the sampled data (Figure 7).

The straightforward estimate of the confidence interval is taken as the percentile of bootstrap estimates beyond a certain threshold (Efron and Tibshirani, 1986). This method is appropriate if the standard deviation of the bootstrap estimates is approximately constant with the mean and the distribution is roughly Gaussian after a monotonic transformation. Conveniently, it is not necessary to compute this transformation—the confidence interval can be inferred directly by counting occurrences. As mentioned before, there may be a bias associated with the number of records used in the estimate. Since the stratification process significantly reduces the data record length, we suggest normalizing the bootstrap estimates by the difference between the mean of the bootstrap estimates and the value of the full record length estimate.

Conclusions

In support of a systematic method of model input selection for finite data and irrelevant input rejection, we introduce a method for determining the optimal binning of histogram-based estimates of mutual information. This optimal binning is adapted to a stratification interval from which independent and identically distributed samples are drawn for bootstrapped estimates of MI which in turn are used to develop confidence intervals for the MI estimate. The methods yield consistent results with actual data from a diesel engine test.

References

Bell, C.B., 1962, "Mutual information and maximal correlation as measures of dependence", *Annals of Mathematical Statistics,* Vol. 33, pp. 587-595.

Cover, T.M. and Thomas, J.A., 1991, *Elements of Information* Theory, Wiley, New York.

Deignan, P. B., Jr., King, G. B., and Meckl, P. H., 2003, "Combinatorial Optimization of Mutual Information Estimates Using Time-Delayed Input Coordinates," *Intelligent Engineering Systems Through Artificial Neural Networks*, Vol. 13, pp. 267-272, ASME Press, New York.

Efron, B., 1979, "Bootstrap Methods: Another Look at the Jackknife", *The Annals of Statistics*, Vol. 7, No. 1, pp. 1-26

Efron, B. and Tibshirani, R., 1986, "Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy", *Statistical Science*, Vol. 1, No. 1, pp. 54-75.

Hall, P. and Morton, S.C., 1993, "On the Estimation of Entropy", *Annals of the Institute of Statistical Mathematics*, Vol. 45, No. 1, pp. 69-88.

Shao, J. and Tu, D., 1995, *The Jackknife and the Bootstrap*, Springer-Verlag, New York.

Scott, D.W., 1992, *Multivariate Density Estimation: Theory, Practice, and Visualization*, Wiley, New York.

Treves, A. and Panzeri, P., 1995, "The Upward Bias in Measures of Information Derived from Limited Data Samples", *Neural Computation*, Vol. 7, pp. 399-407.

Acknowledgments

The authors gratefully acknowledge the support of Cummins Inc. in making this research possible.